DSA5101 Python Machine Learning Project

Li Xiaoli

Data Set

- For our group project, you will use Bank Marketing data **bank-full.csv** to predict whether a client will subscribe a term deposit.
- You are required to use **5-fold cross validation settings (to build your training and test set).**
- **You should implement** at least 3 different machine learning (classification) algorithms, and compare and report their performance.
- Note if you want to perform oversampling and feature selection, you should only do them on the training set. You should not do oversampling and feature selection on the whole data sets, as they will generate over-optimistic performance.

Problem Statement and Performance Evaluation

- The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.
- **MCC** should be used for performance evaluation as it can evaluate well even for imbalanced case.

Matthews correlation coefficient

- MCC is a measure of the quality of binary (two-class) classifications, introduced by biochemist Brian W. Matthews in 1975

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

- If any of the *four sums* in the denominator is 0, then MCC=0. MCC is useful when the two classes are of very different sizes

Matthews, B. W. (1975). "Comparison of the predicted and observed secondary structure of T4 phage lysozyme". Biochimica et Biophysica Acta (BBA) - Protein Structure. ***405*** *(2): 442–451.*

	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

Group & Submission

- **Group:** Each group has 2-3 students – forming by yourself.
- **Submission**
  - Zip your slides and well-documented codes and upload to Luminus\Files\Python ML Projects_Submission by **Sep 26**
  - – Slides
  - – Well-documented codes (including packages that need to install).
  - – Audio/video presentation - you record presentation audio/video.

Suggested Project Slides

- Simple Data Set and Problem Statement Description.
- Dataset Pre-processing (including data exploration & visualization, feature engineering, feature selection, data cleaning etc.) .
- Experimental Study and Analysis.
- Summary of Project Achievements (including the insights from the project, e.g. feature importance analysis, how to use the prediction results for business).
- Future Directions for further Improvements.

Thank You

Contact: xlli@i2r.a-star.edu.sg if you have questions